DOCUMENT RESUME

ED 403 313 TM 026 046

AUTHOR Stancavage, Fran; And Others
TITLE Study of Exclusion and Assessibility of Students with Disabilities in the 1994 Trial State Assessment (TSA) of the National Assessment of Educational Progress (NAEP).
SPONS AGENCY National Academy of Education, Washington, D.C.; Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Apr 96
CONTRACT R999A9001-90
NOTE 51p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Academic Achievement; *Data Collection; *Disabilities; *Educational Assessment; Grade 4; Intermediate Grades; *Limited English Speaking; National Surveys; Reading Tests; Sample Size; *Sampling; Special Education; Student Evaluation; Test Results; *Trend Analysis
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

The National Assessment of Educational Progress (NAEP), a survey of national trends in educational achievement, is attempting to expand its inclusion of students with disabilities or limitations that have previously caused them to be excluded from the assessment. The study described was a precursor to the 1996 changes in NAEP inclusion procedures. It was conducted in conjunction with the 1994 State NAEP fourth grade reading assessment. Study questions about exclusion and assessibility were addressed by collecting independent information on a four state sample of students selected for the 1994 fourth grade reading Trial State Assessment (TSA) who had also been identified by their local schools as having the Individualized Education Plans of special education. The final sample involved 416 students from 123 schools. Of these students, 185 had been excluded from the TSA and 231 were designated as "included," even though a few were absent on the day the TSA was administered. Results suggest that a high proportion of students with disabilities can and should participate in the NAEP reading assessment, but the current NAEP instrument does not provide sufficient accuracy in the lower achievement ranges to support disaggregated reports of proficiency levels for these students. If the goal is to achieve a level of measurement that would allow information about these students to contribute to the estimates of states' overall performances, the large majority of fourth graders are assessable on the current instrument. Suggestions are given for modifying the NAEP instrument. Appendixes discuss calculating the criterion for assessibility and regression functions for prediction of scores. (Contains 10 figures and 21 tables.) (SLD)

ED 403 313

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
FRAN STANCAVAGE

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

Study of Exclusion and Assessibility of Students with Disabilities in the 1994 Trial State Assessment (TSA) of the National Assessment of Educational Progress (NAEP)[1]

Fran Stancavage
Don McLaughlin
Robert Vergun
Cathy Godlewski
Jill Allen

John C. Flanagan Research Center
American Institutes for Research
Palo Alto, CA.

TM026046

[1]Paper presented at the annual meeting of the American Educational Research Association, New York, April, 1996. The work reported herein was done under the auspices of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment and was supported by Grant No. R999A9001-90, administered by the Office of Educational Research and Improvement, U.S. Department of Education.

BEST COPY AVAILABLE

Introduction

Recent educational trends, including the passage of the *Goals 2000: Educate America Act*—which calls for academic standards and assessments that are meaningful, challenging, and appropriate for all students, have prompted many large scale assessment programs to reevaluate their guidelines and procedures for assessing students with disabilities. The National Assessment of Educational Progress (NAEP), an important indicator of trends in national educational achievement since 1969, is among those seeking to expand inclusion. NAEP is a Congressionally-mandated survey of student achievement that assesses proficiency in core subject areas for national samples of students at grades 4, 8, and 12. Since 1990, NAEP, which is conducted by the National Center for Education Statistics (NCES) under the policy guidance of the National Assessment Governing Board (NAGB), has also reported state-level results for states participating in the voluntary State NAEP assessment program.

Currently, NCES and NAGB, working with the NAEP contractors, have begun a number of changes intended to increase inclusion of students with disabilities as well as those with limited English proficiency (LEP students). Using a split-sample design that will assure NAEP's capacity to maintain achievement trends while the new procedures are being implemented, assessment accommodations and bilingual Spanish-English assessment materials were made available to samples of students in the recently completed 1996 mathematics and science assessments. In addition, assessment administrators for the majority of schools in both the national and state NAEP programs operated under new inclusion guidelines in 1996 that are expected to decrease the frequency of exclusions.

The study reported here was one precursor to these 1996 changes in NAEP inclusion procedures. It was conducted in conjunction with the 1994 State NAEP fourth-grade reading assessment, an assessment in which students were asked to read fairly lengthy texts drawn from actual books and stories similar to those they might encounter in class, and to demonstrate their understanding of what they had read by answering a series of multiple-choice and constructed response questions; the format of the latter required students to write their answers in their own words.[2]

Carried out under the auspices of the National Academy of Education (NAE) Panel on the Evaluation of the NAEP Trial State Assessment (TSA)[3], our study was intended to help NCES

---

[2]For a more complete description of the 1994 NAEP reading assessment, see *NAEP Reading Consensus Project, Reading Framework for the 1992 and 1994 National Assessment of Educational Progress* (Washington D.C.: National Assessment Governing Board, 1993). Sample reading tasks from the assessment are included in J.R. Campbell, P.L. Donahue, C.M. Reese, and G.W. Phillips, *NAEP 1994 Reading Report Card for the Nation and the States: Findings from the National Assessment of Educational Progress and Trial State Assessment* (Washington D.C.: National Center for Education Statistics, January 1996).

[3]From 1990 through 1994, State NAEP assessments were authorized on a trial basis only. Beginning in 1996, the trial status was dropped, although the State assessements remain "developmental."

plan for greater inclusion by providing information about the implementation of the then-current exclusion procedures and about the assessibility of students with disabilities in the population sampled by NAEP. In addition, the study informed the Congressionally-mandated NAE evaluation of the TSA by examining the extent to which the accuracy and fairness of State NAEP data might be affected by state-to-state differences in how the exclusion guidelines were applied.

Throughout this paper we use the nominclature of the 1994 NAEP and refer to students with disabilities as Individualized Education Plan (IEP) students.

Research Questions

The 1994 IEP exclusion and assessibility study addressed four primary research questions:

1. What is the assessibility of the excluded IEP students on the NAEP reading assessment used in 1994?

2. What accommodations, if any, would have been needed to include additional students?

3. How was the exclusion decision process for IEP students implemented in the 1994 TSA?

4. Was the exclusion process implemented in a comparable manner in different states? More specifically, were there reading levels at which a child was likely to be included in some states but excluded in others?

Methods

The study questions were addressed by collecting independent information on a sample of students selected for the 1994 fourth-grade reading TSA who also had been identified by their local school personnel as having Individualized Education Plans. Some of these students had actually participated in the TSA, while others had been excluded on the basis of their disability. For the current study, site visitors met with the students to obtain structured measures of reading proficiency; they also interviewed the students' teachers and local NAEP assessment administrators to gather additional information about the students and, more generally, about the implementation of the exclusion process.

Sample

As noted above, the sample for this study was based on students who had previously been sampled for the 1994 TSA. Sampling for the TSA began with the selection of stratified random

samples of approximately 100 public schools and 15 nonpublic schools per state.[4] Each participating school prepared a list of all students at the target grade level, and the NAEP contractor selected from this list a random sample of students ($n \approx 30$) and returned it to the school; the school, in turn, reviewed the sample to identify any withdrawn, out-of-grade, or otherwise ineligible students. At the same time, the school also identified any IEP or LEP students on the sample list and determined whether or not these identified students should be excluded from the TSA.

Forty-one states participated in the 1994 TSA.[5] In order to obtain sufficient statistical power to address the fourth research question, concerning comparability of exclusion decisions across states, we concentrated the study sample into four states. Within each of these states, we began with a simple random sample of schools that participated in the TSA. Within sampled schools, we then attempted to collect information on every IEP student identified on the TSA-sample list, so far as possible within the constraints of a one-day site visit.

State sample. The four states in the sample were selected to represent high and low identification rates and high and low conditional exclusion rates for IEP students. That is, states were sorted into the IEP identification-by-exclusion matrix shown in Figure 1, using participation data from the 1992 TSA in fourth-grade reading. One state was then sampled from each of the four corner cells of the matrix.[6]

---

[4]The sample was restricted to "regular" public and nonpublic schools and did not include special education schools. See J. Mazzeo, N.L. Allen, and D.L. Kline, Technical Report of the NAEP 1994 Trial State Assessment Program in Reading (Washington D.C.: National Center for Education Statistics, December 1995).

[5]Three other jurisdictions, the District of Columbia, Guam, and Department of Defense Overseas Schools, also participated, although the District of Columbia withdrew from the trial after data collection.

[6] The one exception was the high-identified/high-excluded cell, which was empty except for a state that did not participate in the 1994 TSA. A medium-identified/high-excluded state was sampled instead.

Figure 1. IEP Identification-by-Exclusion Matrix

low identification low exclusion	medium identification low exclusion	high identification low exclusion
low identification medium exclusion	medium identification medium exclusion	high identification medium exclusion
low identification high exclusion	medium identification high exclusion	high identification high exclusion

NOTE: For states participating in the 1992 fourth-grade reading TSA, IEP identification rates ranged from 6% to 14% of the total state sample; among IEP-identified students, exclusion rates ranged from 20% to 77%.

School sample. Based on calculations of effect size and the statistical power required to address study question 4, we determined to draw a sample of at least 80 IEP students for each of the four states in the study. However, because we did not know how many IEP students would be in the TSA sample for each school, we began recruitment with a sample of 35 schools per state. Sampling, recruitment, and data collection all overlapped, due to the short time window available for field work, and additional schools were sampled as needed to replace refusing schools or schools with no eligible IEP students.[7] Overall, we contacted 181 schools, of which 34, or 19%, were dropped because they had no eligible IEP students. Seventeen schools (12% of the 147 schools with eligible students) refused to participate or were unable to locate the paperwork that identified students sampled for the TSA; an additional 7 schools were dropped after they had agreed to participate when it became apparent that we would have sufficient student sample sizes without them. Our final sample included 123 schools distributed across the four study states as shown in Table 1.

[7] The TSA is conducted in February. In order to avoid any possible effects on either TSA participation or school exclusion practices (and at the request of the NAEP contractor), we did not begin to recruit schools for our study until after the assessment had concluded. For a number of reasons, however, we wanted to start our recruitment and data collection as soon as possible thereafter. Doing so would decrease uncertainty related to changes in the students' reading ability between data collections; it would also maximize the teachers' and assessment administrators' recall of the exclusion process and assure that data collection could be completed before the end of the school year.

Table 1. School Outcomes by State

	State 1	State 2	State 3	State 4	Total
Sampled	40	44	41	56	181
No Eligible IEP Students	5	9	5	15	34
Refused or No Records	0	6	6	5	17
Dropped - Extra	2	1	1	3	7
Visited	33	28	29	33	123

NOTE: The high rate of sampled schools with no eligible IEP students in State 4 was a consequence of the state's policy of clustering IEP students in selected schools within districts.

Student sample. As indicated above, our design called for us to collect data on all eligible IEP students in the sampled schools, to the extent feasible given our one-day site visits. Because most schools only had one to three eligible students, these time constraints were not a problem except to the extent that they prevented us from making return visits to recover data for absent students or students who had not submitted parental approval slips in time for the data collection.[8] For schools with larger numbers of eligible children, site visitors were instructed to collect data on as many students as possible, alternating between included and excluded students.[9] In some cases it was possible to complete teacher interviews by telephone which had been missed during the site visits. This was not possible for the student assessments, which required a face-to-face interaction.

The TSA sample lists in the 123 schools visited included 480 students whose TSA paperwork indicated that they had been classified as IEPs. Among these 480, 36 (8%) were found to be ineligible for our study because they had withdrawn from the sampled schools before or after the TSA (n=16), because their IEP was based on gifted status rather than disability (n=9), or

[8] We followed the regulations of the individual schools with regard to parental approval. Most schools only required "passive" approval from the parents of selected students; that is, notice of the impending site visit was sent home, and parents who did not wish their children to participate had to notify the school. Some schools, however, required "active" approval, meaning that parents had to return a signed consent form before their children could participate. As would be expected, the majority of parent "refusals" were the consequence of unsigned consent forms at schools requiring active approval.

[9] Within the included or excluded category, site visitors were to interview students in the order in which they appeared on the TSA sample list.

because they were not actually IEP-classified (n=6) or TSA-eligible (n=5). The last category included, for example, students who were not really in the fourth grade. Among the 444 eligible students, student and/or teacher data were collected for 416 (94%). These included 3 students with student data only and 56 students with teacher data only.[10] Student outcomes by state are shown in Table 2.

Table 2. Student Outcomes by State

	State 1	State 2	State 3	State 4	Total
Initial sample	103	134	136	107	480
Not eligible	6	5	12	13	36
Eligible, no data collected	1	10	11	6	28
Student and/or teacher data	96	119	113	88	416

Data Collection

Data was collected by site visitors hired and trained by AIR. All were experienced teachers or researchers who had worked with fourth-grade students and/or students with disabilities. The one-day training session included practice administering the student assessment to special education students recruited from a local school. Data collection included three components:

1) a 10-45 minute assessment for each student,

2) an interview with the teacher or staff person most familiar with the functional reading level and assessibility of the student (5-10 minutes per student), and

3) a brief interview with the teacher or staff person who served as the NAEP assessment administrator (AA) for the school. The latter interview focused on the AA's understanding of the NAEP exclusion process and suggestions for improving that process.

Student protocol. The student assessment was designed primarily to obtain an

[10] Reasons for not obtaining student data included parent refusal (40%), student absence (24%), or severe disability (6%). The remainder (30%) were at schools that had too many eligible IEP students for a one-day site visit. Teacher data were obtained for all of the severely disabled students, all but one of the absent students, and 71% of the students with parent refusal.

independent estimate of reading ability that could inform the study questions on the assessibility of the IEP students and the comparability of exclusion decisions across states. For this purpose we developed an adaptive protocol that could accommodate students over a wide range of ability and would not require pre-screening for assessibility by school staff. The protocol was administered to each student individually, in part because of the nature of the instruments and in part because a number of schools had only one eligible student (who would, by necessity, take the assessment individually) and we wanted to maintain comparable conditions of administration for all students.

The student protocol began with a brief rapport building activity that also allowed site visitors to screen for a minimum level of cooperation and oral English fluency. This was followed by administration of the Woodcock Johnson Broad Reading Cluster (WJ BRC), an adaptive assessment tool that is used extensively with special education populations and covers reading functionality from age 2 through adult. The WJ BRC was field scored by the site visitor, and the assessment was terminated for students who performed below a preestablished criterion.[11] For the remaining students, the assessment continued with the administration of an abridged, and relatively easy, item block from the 1992 TSA titled *Amanda Clement: The Umpire in a Skirt*. The abridged item block included three multiple choice and two constructed response items. As in a regular NAEP administration, the student was required to answer the questions in writing; the site visitors, however, were provided with standard prompts that they could use to encourage students who failed to engage with the assessment task. The constructed response questions were subsequently scored for the study by the Educational Testing Service (ETS), the NAEP contractor, using standard NAEP protocols. To minimize score drift, study papers were mixed with a representative set of student responses from the regular 1992 administration of the *Amanda* prompt. The WJ BRC data provided a standardized measure of reading proficiency that served as the primary basis for answering the research questions on assessibility and cross-state comparability. The NAEP item set was administered provide a more qualitative sense of how these students would perform on NAEP-like tasks.

Teacher protocol. For each student, a structured interview was conducted with a teacher or other staff person familiar with the student's disability status, level of functioning, and school program. The interview also included questions on the kinds of testing accommodations, if any, that the teacher-respondent deemed appropriate for the student and, if the teacher had been involved in the exclusion decision process, reasons for including or excluding the student and general impressions about the NAEP exclusion guidelines. Where ever possible, the interview was done with the same person who had completed the NAEP *IEP/LEP Student Questionnaire* at

---

[11] The WJ BRC consists of two subtests: (1) a letter-word identification test in which the student is asked to read out loud progressively more difficult words, and (2) a close test of passage comprehension to which the student also responds orally. At the lowest levels, students respond to single letters or pictures rather than words or sentences; the assessment ends when the student misses the answers to six questions in a row. Grade equivalent score conversions for each subtest are provided by the publisher, and, based on the judgement of our reading consultants, a grade equivalent below 2.6 on either subtest was preestablished as the criterion for terminating the assessment.

the time of the NAEP administration.[12] Data from the teacher interviews were used primarily in answering the research questions on accommodations and the exclusion process.

Assessment administrator protocol. At each school one teacher or other staff person was designated as the local AA for NAEP. This individual, who was trained by the NAEP contractor, was responsible for finalizing the student sample, administering the TSA assessment, distributing and collecting the various NAEP questionnaires (including the *IEP/LEP Student Questionnaire*), and completing all related paperwork. As such, this person was the gatekeeper for the exclusion decision processes and, in many cases, was actively involved in the exclusion decisions for individual students. A brief structured interview was conducted with the AA at each school that contributed students to our final study sample; the interview covered the respondent's general impression of the NAEP exclusion guideline, and, where applicable, the reasons for including or excluding each IEP student. Data from the AA interview were used to answer research question 3, on the exclusion process.

Results

The results section begins with a brief presentation of selected characteristics of the student, teacher, and AA samples. This is followed by a discussion of results organized around the four research questions listed in the introduction.

Sample Characteristics

Students. Except for research question 4, concerning comparability across states, the state samples were pooled for analysis. The final four-state sample included 416 IEP students in 123 schools. One-hundred-eighty-five of these students had been excluded from the TSA, the remaining 231 students were designated as "included," although a small percentage of the latter had been absent on the day the TSA was administered (and were not subsequently assessed in TSA make-up sessions). Tables 3 through 5 show the demographic characteristics of the sampled students and present information on their disabilities and educational programs. All information in these tables is based on responses to the teacher questionnaires.

As can be seen, the student sample was two-thirds male and two-thirds white (Table 3). Males and non-whites were slightly over represented among the excluded students, but the differences were small and, in the case of gender, not statistically significant. Learning disabilities were given as the primary disability for two-thirds of the students in the sample, while nearly 90% were classified as mildly or moderately disabled (Table 4). Students whose disabilities were described as less severe were somewhat less likely to have been excluded, but even among those classified as only mildly disabled, nearly 40% had been excluded from the

[12]This questionnaire, which covered many of the same topics as our teacher interview—although in less detail—was completed for every IEP or LEP student in the TSA sample. Instructions for the questionnaire indicated that it should be completed by the student's teacher or other knowledgeable staff member.

TSA. With regard to their educational programs, excluded students spent substantially less time mainstreamed.[13] They also were considerably less like to receive reading instruction in a mainstreamed class and considerably more likely to receive all or some of their reading instruction in a special education class (Table 5).

Table 3. Demographic Characteristics of the Student Sample

	Included in TSA N = 231	Excluded from TSA N = 185	All Students N = 416
Gender:			
Male	60%	70%	64%
Female	40%	30%	36%
Race/Ethnicity:			
White	69%	58%	64%
Black	13%	28%	20%
Hispanic	13%	12%	13%
Other	4%	2%	3%

[13]Mainstreaming is generally interpreted to mean that the student is served in the regular classroom and receives at least some of the same instructional content as is offered to other students.

Table 4. Type and Severity of Disability among the Student Sample

	Included in TSA N=231	Excluded from TSA N=185	All Students N=416
Primary Disability			
Mental Retardation	3%	18%	10%
Learning Disability	65%	70%	67%
Sensory Disability	0%	2%	1%
Emotional Disturbance	6%	5%	6%
Speech Impairment	24%	1%	14%
Other Disability	2%	3%	2%
Severity of Disability			
Mild	60%	38%	50%
Moderate	32%	43%	37%
Severe	7%	15%	10%
Profound	1%	4%	2%

Table 5. Characteristics of Educational Programs for Students in the Sample

	Included in TSA N=231	Excluded from TSA N=185	All Students N=416
Median percent time mainstreamed	90%	40%	73%
Reading instruction in mainstream class	69%	26%	50%
Reading instruction in special education class	57%	91%	73%

Teachers. At the time of the TSA administration, the school AA collected *IEP/LEP Student Questionnaires* for each of the IEP students on the TSA sample list. AA instructions indicated that this questionnaire should be completed by someone "who knows the student well." For our study, we attempted to interview the same respondent. As can be seen in Table 6, we were successful in this effort for more than 80% of the cases. Job positions of the "teacher" respondents are also shown in Table 6; nearly all were either the students' special education teachers or regular classroom teachers. In some cases, teachers provided responses for multiple students. The data in Table 6 are presented weighted by the number of students for whom the teacher provided information.

Table 6. School Positions of "Teacher" Respondents

	Same Respondent Who Completed IEP/LEP Questionnaire N (of stdnts) = 339	Not Respondent Who Completed IEP/LEP Questionnaire N (of stdnts) = 67*	All "Teacher" Respondents N (of stdnts) = 406
Special Education Teacher/ Case Manager	65%	54%	64%
Regular Classroom Teacher	28%	30%	28%
Other Teacher	4%	11%	5%
Other	3%	5%	3%

*Includes 7 cases with questionnaire respondent status unknown

Assessment Administrators. Unlike the "teacher" respondents, the AA respondents were much more likely to be non-teaching personnel. Table 7 shows that, weighted by the number of students they represented, the largest proportion of AA responses (39%) were given by school principals or vice principals, followed by counselors or school psychologists (30%).

Table 7. School Position of Assessment Administrator Respondents

	N (of students) = 404
Principal/Vice Principal	38%
Counselor/School Psychologist	30%
Classroom teachers	9%
Others (primarily school resource personnel)	23%

We now turn to a discussion of the results pertinent to each of the research questions posed at the beginning of this paper.

Research Question 1: What is the assessibility of the excluded IEP students on the NAEP reading assessment used in 1994?

Defining assessibility. In considering assessibility, our primary criterion was whether or not the student's reading proficiency was in a range that could be measured, at least crudely, by the current NAEP instrument. We were interested in a level of measurement that would allow information about the student's achievement to contribute to estimates of a state's overall performance; a more stringent criterion might be needed if the goal were to characterize the performance of IEP students separately.

As discussed in the methods section, the WJ BRC score provided an independent measure of reading proficiency based on an adaptive instrument that accommodated a wide range of student capabilities. For our sample, it correlated reasonably well (r=.75) with NAEP reading scores (plausible values).[14] We therefore addressed the question of assessibility by further examining the relationship between the WJ BRC scores and the NAEP plausible values of included IEP students in our sample. In particular, we were interested in determining whether this relationship changed for different levels of reading proficiency (as measured by the WJ BRC) and, more specifically, whether there was a WJ BRC score level below which the NAEP plausible values no longer captured any of the observed variation in the proficiency distribution.

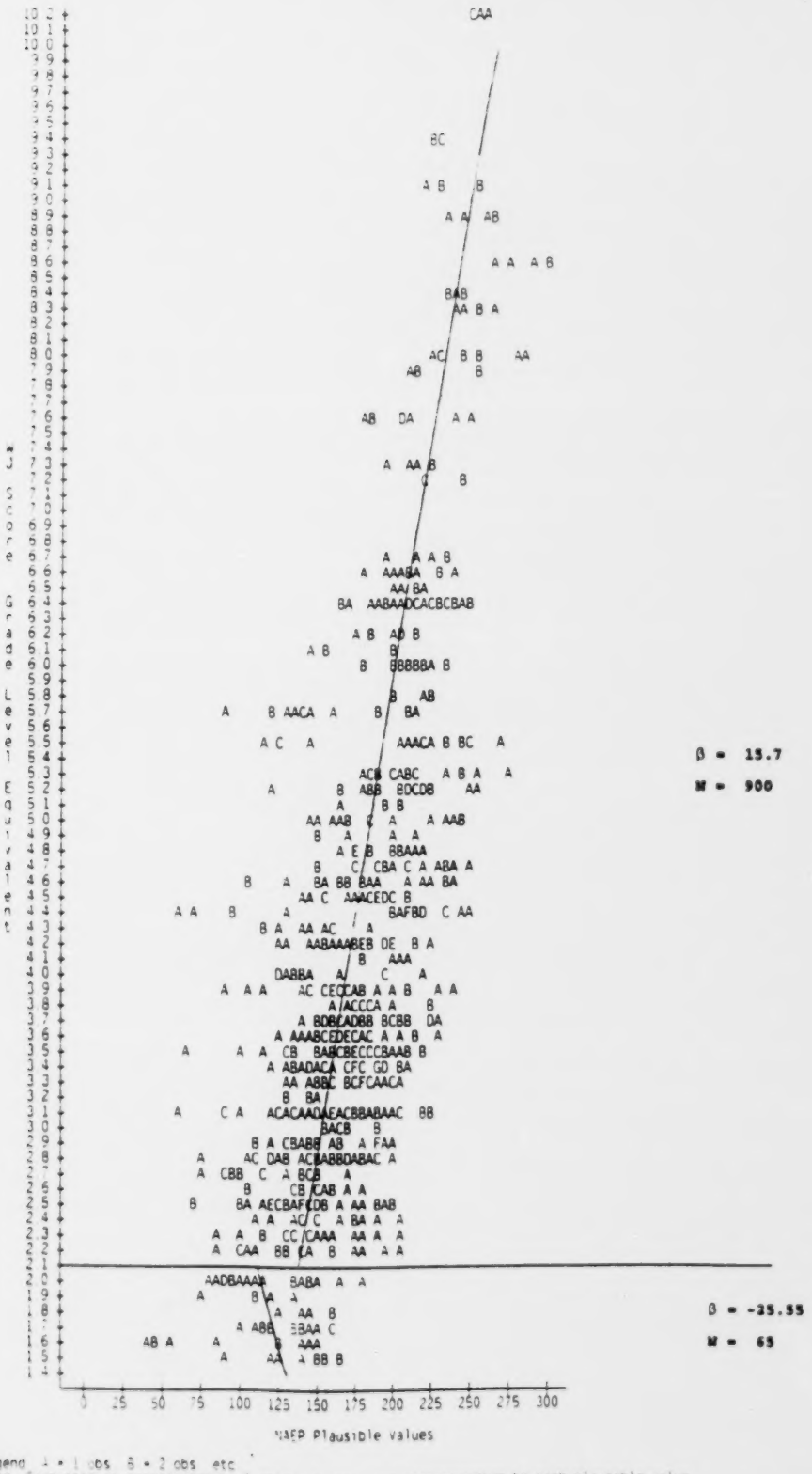
The final results, shown in Figure 2, suggest that meaningful performance on NAEP is possible for fourth-grade students with grade-equivalent WJ BRC scores at or above 2.1. For purposes of the present analysis, we used this cut point as the criterion for assessibility. However, because the 2.1 cut point falls very close to the bottom of the proficiency range for which comparative data are available (i.e., very near the bottom of the distribution for IEP students who were included in the 1994 TSA), the criterion should be interpreted as approximate rather than precise. For further discussion of the method used to estimate the cut point, see Appendix A.

---

[14]NAEP assessment results are based on a matrix sampling model in which different students respond to different assessment tasks, and results are combined on an overall proficiency scale using Item Response Theory (IRT). The design is intended to provide reliable proficiency estimates for *groups* of students; because of the relatively small number of assessment tasks attempted by any one student, student-level proficiency estimates are less precise. For each student, NAEP therefore computes a distribution of so-called "plausible values" and retains five values, sampled from this distribution, to represent equally plausible estimates of that student's achievement. For further details about NAEP scaling and analysis, see the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*, J. Mazzeo, N.L. Allen, and D.L. Kline, op cit.

Figure 2.

Plot of Woodcock Johnson Broad Reading Cluster Scores by NAEP Plausible Values



Legend: A = 1 obs, B = 2 obs, etc.
Note: Each case is represented by 5 observations, one corresponding to each plausible value.

BEST COPY AVAILABLE

Proportion of students judged assessable. Having thus established a measure of assessibility, we next looked to see how the included and excluded students distributed on the WJ BRC measure and what proportion of the students in our total sample of included and excluded IEP students were assessable. We noted first that the median WJ BRC grade-equivalent score for both included and excluded students were substantially higher than 2.1 (As shown in Figure 3, the median was 2.7 for excluded students and 3.7 for included students, but the distributions overlap substantially.) More precisely, as shown in Table 8, *fully 83% of the students in our sample had scores at or above 2.1 and thus met our criterion for assessibility. The assessable group included 70% of the excluded students and 93% of those who had participated in the TSA.* These figures are very consistent with the National Center on Educational Outcomes' 1994 estimate that approximately 85% of students with disabilities could be included in large-scale assessments, with the majority not requiring any major accommodations.[15]

[15]National Center on Educational Outcomes, *Synthesis Report 15: Recommendations for Making Decisions about the Participation of Students with Disabilities in Statewide Assessment Programs* (Minneapolis, MN: Author, 1994), p. 5, and private communication, Kevin McGrew, NCEO Senior Researcher, May 23, 1995.

Figure 3.

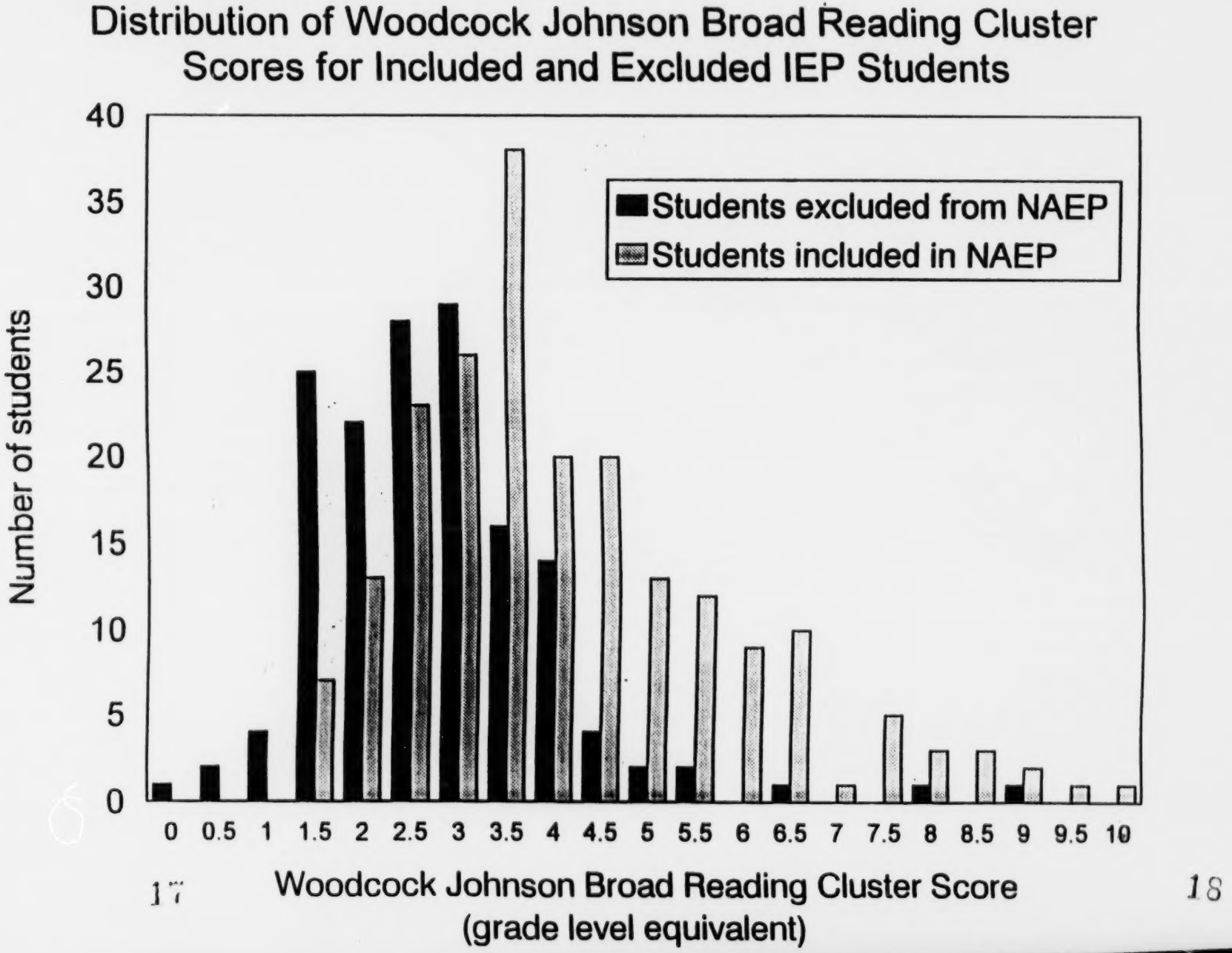

Distribution of Woodcock Johnson Broad Reading Cluster Scores for Included and Excluded IEP Students
Students excluded from NAEP
Students included in NAEP
Number of students
0
5
10
15
20
25
30
35
40
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5 5.5 6 6.5 7 7.5 8 8.5 9 9.5 10
Woodcock Johnson Broad Reading Cluster Score
(grade level equivalent)

Table 8. Percent of Included and Excluded Students Predicted to be Assessable on NAEP (Based on Woodcock-Johnson Broad Reading Cluster Score)

	Included in TSA N=206	Excluded from TSA N=158	All Students N=364
Predicted Assessable on NAEP*	93%	70%	83%
Predicted Not Assessable on NAEP**	7%	30%	17%

* Woodcock-Johnson Broad Reading Cluster score at or above 2.1 grade level equivalent

** The four students who were not assessed during our site visits due to severe cognitive impairment were also counted as not assessable on NAEP.

Performance on the abridged NAEP item set. A second perspective on the assessibility of students in our sample is obtained by reviewing performance on the full individual student assessment administered for our study. This performance is summarized in Table 9. Recall that the assessment was designed to be incremental and to terminate when further assessment might be stressful or inappropriate to the student's reading level. Therefore, prior to the site visits and based on the advice of our expert advisors, we had established an *a priori* cut score on the WJ BRC below which the assessment was to be terminated without attempting the NAEP item set. As it turned out, this criterion for termination (grade-equivalent score below 2.6 on either subtest of the WJ BRC) was substantially more stringent than the criterion we later derived for assessibility. Consequently, many students whom we later felt might have been able to meaningfully attempt the NAEP item set were not given this opportunity. Nearly a quarter of our total student sample, including 17% of the included students and 33% of the excluded students, fell into this category; that is, they were predicted assessable based on the 2.1 grade-equivalent cutscore described above, but they were screened out of the site visit protocol before the NAEP item set.

Table 9. Student Assessment Results

	Included in TSA N = 206	Excluded from TSA N = 158	All Students N = 364
Not possible to assess due to severity of disability	0%	3%	1%
Began assessment, but screened out in warm-up activity	0%	1%	<1%
Attempted Woodcock-Johnson Broad Reading Cluster but screened out before NAEP item set *	24%	58%	39%
Attempted NAEP item set, but answered no questions correctly	10%	9%	10%
Answered at least one multiple choice item correctly, but did not answer either of the constructed response items correctly	23%	19%	21%
Answered at least one constructed response item correctly **	43%	9%	29%

* Screened out based on the criterion of one or both Woodcock-Johnson Broad Reading Cluster subscale scores < 2.6 (grade equivalent).

** All but 11 of the 104 students who answered at least one constructed response correctly also answered at least one multiple choice item correctly.

Among the students who did attempt the NAEP item set, we distinguished between three levels of performance: students who answered none of the five items correctly; those who answered at least one of the three multiple-choice items correctly, but neither of the constructed-response items; and those who were successful on one or both of the constructed-response items. While a five-item reading measure is not highly reliable, when we compare included and excluded students on this measure, we do see that the proportion of excluded students who reached the top category is quite small; only 9% as compared to 43% of the included students.

The explication of this performance difference between included and excluded students becomes clearer when we examine the relationship between performance on the NAEP item set and the students' WJ BRC scores. For the three categories of performance that we considered, there was substantial overlap in WJ BRC scores, but, on average, the WJ BRC scores tended to

be higher for students whose performance on the NAEP item set was also classified higher. (See Table 10.) Further, among those who performed in the highest category (i.e., answered at least one constructed-response item correctly), the median WJ BRC score was 4.7, and the middle 50 percent of students had scores that ranged from 3.6 to 6.3. These are rather high reading levels, and relatively few of the excluded students in our sample fell into this range.

Table 10. Relationship Between Performance on the NAEP Item Set and WJ BRC Scores

	Included in TSA N=157	Excluded from TSA N=60	All Students N=217
	WJ BRC scores: median (Q1-Q3)*	WJ BRC scores: median (Q1-Q3)	WJ BRC scores: median (Q1-Q3)
Answered no items correctly	3.5 (3.0-4.4) N=20	3.4 (2.9-3.8) N=15	3.4 (3.0-4.2) N=35
Answered at least one multiple choice item correctly, but neither of the constructed response items	3.9 (3.5-4.5) N=48	3.3 (3.2-3.9) N=30	3.7 (3.2-4.4) N=78
Answered at least one constructed response item correctly	4.8 (3.7-6.4) N=89	4.2 (3.1-5.0) N=15	4.7 (3.6-6.3) N=104

*Q1-Q3 gives the highest and lowest WJ BRC score values for the middle 50 percent of students (students in the second and third quantiles) in each table cell.

NOTE: As a consequence of the protocol used for the individual student assessments, these data only include students with grade-equivalent scores of 2.6 or higher on each subtest of the WJ BRC.

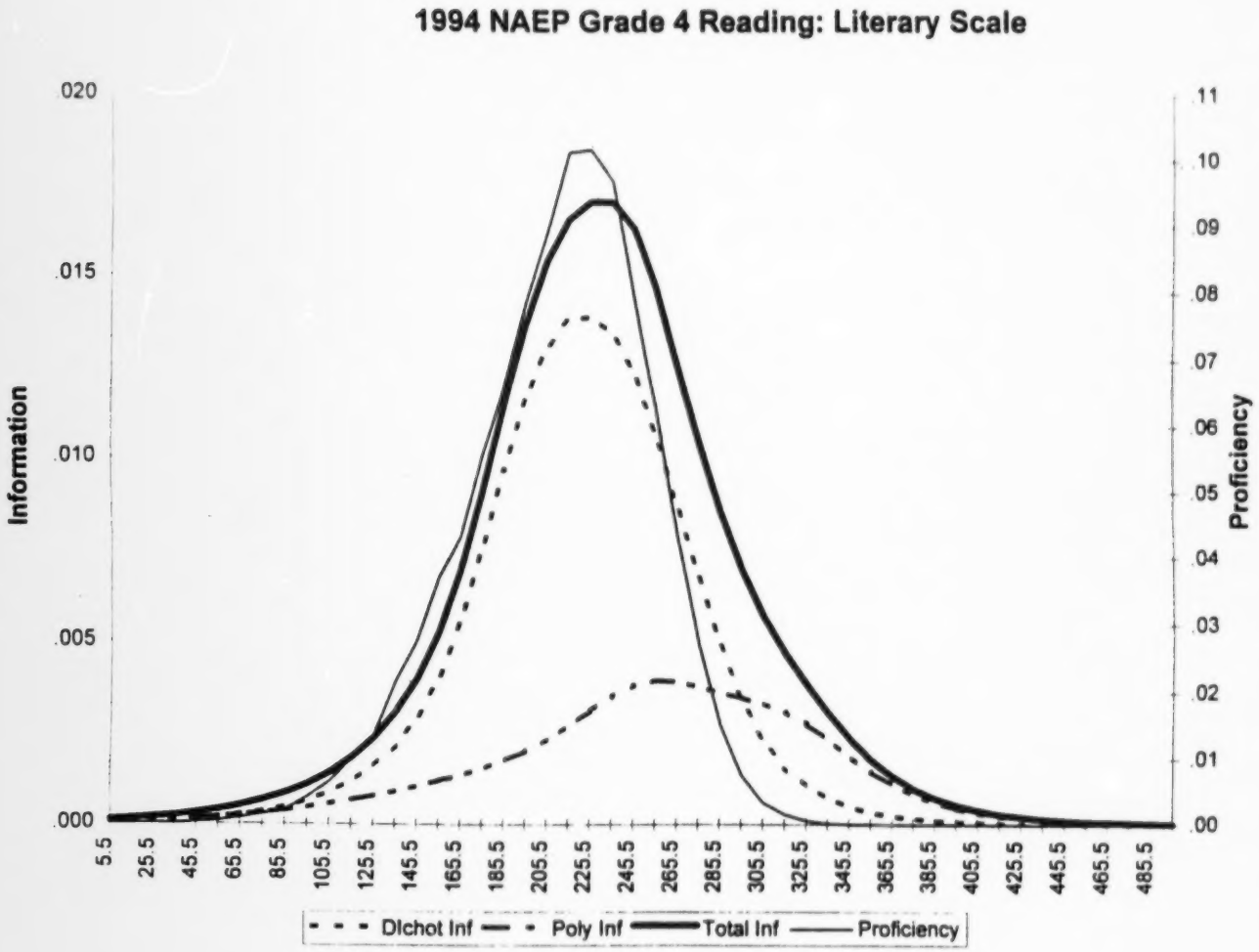
The constructed-response items, of course, require that students not only understand what they have read, but also be able to express this understanding in writing. Because different students may have had different opportunities to practice their writing skills in school, it is possible that the observed association between success on these items and higher WJ BRC reading scores may reflect differences in the learning opportunities of different categories of special education students as well as pure differences in reading ability. In either case, the data still suggest that

most of the discrimination that the NAEP reading assessment offers for students in the lowest proficiency range derives from the multiple choice item.

Item difficulty and the impact of constructed response items in the 1994 NAEP reading assessment. Further evidence that points in the same direction is given in Figure 4, which shows, for one of the two subscales on the 1994 fourth grade reading assessment, the amount of information provided by the dichotomously-scored and polytomously-scored items at different points on the proficiency scale. The polytomously scored items, which represent one class of constructed response items (primarily the extended response items that require more lengthy answers) are generally providing information for students much higher on the proficiency scale than the dichotomous items.[16]

[16]"The item information function (IIF) expresses the relationship between the information provided by an item and the proficiency variable. ..information, in this context, is inversely proportional to error of measurement and hence the higher the information the less measurement error in estimating examinees' proficiencies from the item. The information function provides the test developer with data about where in the proficiency range the item is providing the best (most precise, or with least error) measurement. ... The test information function (TIF) is simply the sum of the individual item information functions. ... The test information is inversely proportional to the standard error of measurement of the test. Hence the plot of the test information function can be used to determine ranges in proficiency for which the standard error is sufficiently low (where the information is high) for accurate measurement." J.E Carlson, "Information Provided by Polytomous and Dichotomous Items on Certain 1994 NAEP Instruments," paper presented at the annual meeting of the American Educational Research Association, New York, April 1996.

Figure 4. *Student proficiency and test information distribution functions (TIFs) for dichotomously-scored items, polytomously-scored items, and for all items combined*



SOURCE: Unpublished data provided by J.E. Carlson, Educational Testing Service, March 1996

It has been evident for some time that the items in the new NAEP reading assessments, written to the 1991 framework, tend to be concentrated in the mid and higher difficulty ranges relative to the target population and that the extensive use of constructed response items has contributed to this trend. Sixty-three percent of testing time in the 1994 fourth-grade reading assessment is spent on constructed-response items, and the proportion of testing time devoted to these items at the higher grades is even higher, exceeding 75%. Therefore, while our analyses do indicate that the current NAEP assessment can yield rough estimates of reading ability for the large majority of students with disabilities, the test is clearly not pitched at a level that can provide a very fine-grained measure of reading for these students, many of whom are reading a year or more below grade level. *We therefore conclude that, while a high proportion of students with disabilities can and should participate in the NAEP reading assessments, the current NAEP instrument does not provide sufficient accuracy in the lower achievement ranges to support disaggregated reports of proficiency levels for these students.*[17]

[17]Other issues also arise with regard to the value of disaggregated reporting of achievement scores for special education students. These include problems with the definitions of the reporting groups, given that disability labeling is highly impacted by education policy and funding, and given that the population that gets labeled varies from state to state as well as from year to year.

Research Question 2. What accommodations, if any, would be needed to include additional students?

Accommodations recommended by teacher respondents. The 1994 TSA did not permit any testing accommodations for students with special needs As part of our teacher interviews, however, we asked respondents whether they would recommend NAEP participation with accommodation for each of their students in our sample and, if so, what accommodations they would recommend. In addition, we asked what accommodations, if any, were normally provided for these students for participation in state or local assessments.

More specifically, teacher respondents were briefed on the broad characteristics of the NAEP reading assessment and shown a sample item block, then asked whether, in light of the length and format of the assessment, they would affirm or modify the original inclusion/exclusion recommendations for each of their students. That is, they were asked whether, in their judgement, their students "could have taken the assessment," could have taken it with accommodations, or would more appropriately have been excluded. Responses to this query, in which the teachers' choices were unconstrained by the need to justify accommodation or exclusion based on any specific criteria, are shown in Table 11. As can be seen, *the teachers favored accommodations for more than half of their students with disabilities, including nearly two-thirds of those who had been included in the 1994 TSA and over 40% of those who had been excluded.* In most of the remaining cases the teachers reaffirmed the original inclusion/exclusion recommendations that had been made for these students at the time of the TSA assessment.[18]

Table 11. What Teachers Recommended Regarding Inclusion/Exclusion after Reviewing Sample NAEP Reading Block and Having Been Offered Inclusion with Accommodation as an Option

	Included in TSA N = 231	Excluded from TSA N = 185	All Students N = 416
Include without accommodation	31%	5%	20%
Include with accommodation	62%	43%	53%
Exclude	7%	52%	27%

These percentages are consistent with the finding from the 1995 NAEP field test in mathematics that 49% of the grade-four IEP students had accommodations for achievement testing specified in their individualized education plans. Our teachers' propensity to recommend accommodations may also have been related to the fact that these professionals are most used to

[18] Teacher respondents reported having been involved in the original inclusion/exclusion decision for 63% of the students in our study.

thinking in terms of instructional activities, wherein the overriding goal is to facilitate each student's learning. They are less used to considering the advantages of standardization in large scale assessments. Nevertheless, the upshot is that, *had the teachers' unconstrained recommendations been followed, only 20% of the 416 IEP students in our study would have been assessed without accommodations, as opposed to the 56% who actually were assessed.*

We next focused on the *types* of accommodations that the teachers would recommend. Accommodation suggestions for the learning disabled (LD) students who accounted for over two-thirds of our total student sample are presented in Table 12. Suggestions for extended time and/or shorter tests were most ubiquitous, having been proposed for nearly all (85%) of the LD students recommended for inclusion with accommodation. These were followed in frequency by suggestions for oral reading of directions (74%), small group or individual administration (70%), and oral response (56%).

Table 12. Adaptations Most Frequently Suggested for Learning Disabled Students (Among LD Students Recommended for Inclusion with Adaptation)

	Included in TSA N = 108	Excluded from TSA N = 60	All Students N = 168
Presentation Format			
▸ No presentation adaptations	19%	13%	17%
▸ Oral reading of directions *plus* interpretation of directions	38%	40%	39%
▸ Oral reading of directions only	31%	42%	35%
▸ Large print	18%	12%	15%
Response Format			
▸ No response adaptations	29%	18%	25%
▸ Oral response *plus* assistance and interpretation with response	19%	35%	25%
▸ Oral response only	30%	33%	31%
▸ Assistance and interpretation with response only	13%	13%	13%
▸ Mark in test booklet *	18%	17%	17%
Setting of Test			
▸ No setting adaptations	37%	18%	30%
▸ In special education class or small groups	46%	62%	52%
▸ Alone in a test carrel	17%	20%	18%

Table 12. Adaptations Most Frequently Suggested for Learning Disabled Students (Among LD Students Recommended for Inclusion with Adaptation)

	Included in TSA N = 108	Excluded from TSA N = 60	All Students N = 168
Timing of Test			
▸ No timing adaptation	11%	7%	10%
▸ Extended time *plus* shorter version of test	28%	42%	33%
▸ Extended time only	19%	32%	23%
▸ Shorter test only	37%	15%	29%

*Would not be an accommodation; matches response mode of current NAEP.

Impact of accommodations on student performance. If accommodations are to be offered, their impact on performance must be understood before the scores of accommodated students can be combined with the scores of other test takers. Our study was not designed to address this question systematically, and the results provide only limited insight into this problem. We did explore the possibility that the NAEP items used in our study had a different relationship to theta (the latent trait—in this case reading proficiency—assumed by the IRT scale) when we administered them than they had had when they were administered in the regular NAEP assessment.[19]

For this comparison, we used only students who had participated in the actual TSA and for whom we therefore had an estimate of theta (proficiency). We considered what the pass rate for each of the items in the NAEP item block "ought to have been," based on the estimated thetas for the participating students, versus the actual pass rates achieved under the conditions of our study. Unfortunately, the statistical comparisons could only be approximate, because our students were participants in the *1994* TSA, while the items were drawn from a released 1992 item set that had not been readministered in 1994.

The results of the comparison are shown in Table 13. As can be seen, the item difficulties (p-values) were reasonably similar, and the differences were not all in the same direction.

[19]Our study condition can be viewed as an "accommodation" in that the test was considerably shortened and the student worked one-on-one with the site visitor; the latter was instructed to offer repeated encouragements if the student failed to engage with the task or stopped prematurely.

Table 13. NAEP Item Difficulties (p-values)

	As Administered to Study Sample N=149	As Estimated from Students' Thetas N=149
Item # 1 (multiple choice)	.50	.56
Item # 2 (multiple choice)	.56	.53
Item #3 (constructed response)	.46	.41
Item #4 (constructed response)	.29	.25
Item #5 (multiple choice)	.45	.46

NOTE: Estimation based on 1992 item parameters and 1994 theta estimates. Only students who participated in the 1994 TSA are included in the analysis.

Students with orthopedic or sensory disabilities. The above discussion focuses on the learning disabled students, who, as noted, made up more than two thirds of our sample. (Population estimates derived from the NAEP *IEP/LEP Student Questionnaire* show that, across all of the participating jurisdictions in the 1994 TSA, 75% were classified as learning disabled. This percentage varies from state to state, however.) We also looked more closely at the prevalence and assessment requirements of students with orthopedic or sensory disabilities because these students might require more specialized or technologically sophisticated accommodations to participate in NAEP. As it turned out, only 16 out of the 416 students in our sample (4 percent) had such disabilities as either their primary disability or in addition to some other primary disability. This is consonant with population estimates from the *IEP/LEP Student Questionnaire* which indicated that only two percent of the fourth-grade regular school students in the TSA states had primary disabilities that were orthopedic or sensory. An additional 3 percent of this population were classified as having multiple disabilities.

Among the 16 students with orthopedic or sensory disabilities in our sample, four had been included in the 1994 TSA and 12 had been excluded. At the time of our interviews, our teacher respondents would have recommended inclusion without accommodation for only one (a visually impaired student), inclusion with accommodation for five others (one hearing impaired student and four learning disabled students who also had orthopedic or visual impairments), and exclusion for the remainder.[20] Thus it seems clear that the capacity to provide specialized

[20]The 10 students for whom the teacher respondents recommended exclusion comprised 2 students with orthopedic impairments as a primary disability, 2 students with hearing impairment as a primary disability, 2 learning disabled students who also had orthopedic or visual impairments, and 4 mentally retarded students who were also hearing impaired, orthopedically impaired, and/or visually impaired.

accommodations for students with these kinds of physical impairments would have very little impact on the overall proportion of students with disabilities who would be included in NAEP.

Research Question 3. How was the exclusion decision process for IEP students implemented in the 1994 TSA?

As described in an earlier section of this paper, the TSA student sample in each participating school was drawn randomly from a list of all grade-eligible students. Local school personnel were then given the opportunity to exclude any of the students with IEP or equivalent classifications (or any of the students with limited English proficiencies), if they were judged incapable of taking the assessment. The criteria by which local personnel were to make these judgements were given in the "Criteria for Excluding Students from the Assessment," shown in Figure 5. These criteria state that the AA, with the advice of staff members knowledgeable about the students, may exclude IEP students if a) the student is mainstreamed less than 50% of the time in academic subjects and is judged incapable of participating meaningfully in the assessment, or b) the IEP team or equivalent group has determined that the student is incapable of participating meaningfully in the assessment.

Figure 5. Criteria for excluding students from the assessment: 1990 - 1994

The intent is to assess all selected students. Therefore, all selected students who are capable of participating in the assessment should be assessed.

Some of the students identified on the Administration Schedule as Limited English Proficient (LEP) or as having an Individualized Education Plan (IEP) may be incapable of participating meaningfully in the assessment. The Assessment Administrator, with the advice of staff members knowledgeable about the IEP/ LEP students, may exclude such students, as described below.

1. A student identified on the Administration Schedule as LEP may be excluded from the assessment if he/she:

    - Is a native speaker of a language other than English;

    AND

    - Has been enrolled in an English-speaking school (not including a bilingual education program) for less than two years;

    AND

    - Is judged to be incapable of taking part in the assessment.

2. A student identified on the Administration Schedule as having an IEP or equivalent classification may be excluded from the assessment if:

    - The student is mainstreamed less than 50 percent of the time in academic subjects and is judged incapable of participating in the assessment;

    OR

    - The IEP team or equivalent group has determined that the student is incapable of participating meaningfully in the assessment.

3. IEP/LEP students meeting the above criteria should be assessed if, in the judgment of school staff, they are capable of taking the assessment.

WHEN THERE IS DOUBT, INCLUDE THE STUDENT.

One aim of our study was to learn more about how this exclusion process was actually implemented at the local level. To this end, we included in our teacher and AA interviews a number of questions about the participants in the exclusion decision process and about the specific criteria used to determine which students would be excluded and which would not.

Generally, our respondents reported that the exclusion or inclusion decisions were reached in a group process, with several individuals participating. And, as shown in Table 14, both the AAs and those teacher respondents who had also completed the *IEP/LEP Questionnaires* at the time of the TSA assessment tended to report that they had been among the participants. Our two respondents were not, however, always in agreement with one another in reporting which of them had been involved, and this suggests that the collaborative or advice seeking process was more informal than formal in many cases. Generally speaking it seemed that the teacher respondent, who very often worked with the student on a daily basis, was in the best position to judge the student's capacity for meaningfully participating in the TSA. It would have been the AA, however, who would have had ultimate responsibility for completing the paperwork that recorded the assessment outcome for each student in the school sample, as well as for verifying that the correct students were present for the assessment.

Table 14. Reported Role in TSA Inclusion/Exclusion Decision

	Teacher Respondents		AA Respondents N (of students)=404
	NAEP Questionnaire Respondents N (of students)=339	Not NAEP Questionnaire Respondents N (of students)=74*	
Involved in decision	69%	35%	71%
Not involved or does not remember	31%	65%	29%

*Includes 7 cases with questionnaire respondent status unknown

Among those who did report being involved in the decision process, 43 percent of the teacher respondents reported that they had seen a copy of the official NAEP exclusion guidelines when they were making their decisions, an equal number reported that they had not seen these guidelines, and the remaining 14 percent could not remember. (All AAs would necessarily have seen these guidelines because they are included in their manual and covered in the training that they receive from the NAEP contractor.) When they were shown the guidelines in the context of our interview, however, nearly all of these teacher and AA respondents indicated that the guidelines were clear (Table 15). The most frequent reasons for not finding the guidelines clear involved trying to apply the "percent time mainstreamed" criteria in schools that served their students with disabilities in the regular classroom setting and being uncertain how to operationalize the phrase "participate meaningfully in the assessment."

Table 15. Reactions to NAEP Exclusion Guidelines
(Among Those Reporting that They Participated in TSA Inclusion/Exclusion Decision)

	Teacher Respondents N (of stdnts) = 260	AA Respondents N (of stdnts) = 288
Saw copy of guidelines when deciding whether student should participate		
Yes	43%	
No	43%	
Doesn't remember	14%	
Considers guidelines to be clear		
Yes, clear	93%	81%
Mostly clear, but some problems	7%	16%
No, not clear	0%	3%

For each individual student, respondents who had participated in the exclusion process were further asked to identify—from a prespecified list—the one factor that had primarily influenced their decision, as well as other factors that had contributed to the decision. The frequency with which different factors were cited as primary is shown in Table 16. In this table, each student is counted only once, based on the primary factor cited by the teacher respondent. Data from the AA respondent were substituted (where available) only if the teacher respondent had not answered the question. The data show that, as one might expect, the students' reading level was by far the factor most commonly cited as primary to the decision. Most often respondents chose the option that was worded "Reads/doesn't read well enough to take the NAEP," but another sizeable fraction of respondent chose the option "Reading/not reading at grade level," particularly in the cases where the decision was to exclude. The latter seems somewhat further afield from what had been intended by the developers of NAEP, who specify in the *Reading Framework* that the assessment should include reading passages suitable for the least proficient students in the class as well as some appropriate for the most proficient.[21]

After reading level, the next most common factor cited as primary was that the IEP

[21]NAEP Reading Consensus Project, op cit., 20. The *Framework* further indicates that the passages for the least proficient should be pitched a couple of grades below the target grade (e.g., at grade two for a fourth-grade class). However, as noted in the discussion of the first research question, the actual 1994 fourth-grade reading assessment was skewed toward passages and items that were relatively difficult for the majority of students, let along those who were least proficient.

specifies that the student should or should not participate. This appears congruent with the exclusion criteria, which contain the option that a student be excluded if the IEP team or equivalent group has determined that the student is incapable of participating meaningfully in the assessment. The same cannot be said for the factors driving the assessment decisions for the 12% of included students who were reportedly tested because they had previously been included in their own state or district assessment. All of these responses, however, suggests the respondents' struggles to come up with a more quantifiable decision rule than simply "participate meaningfully in the assessment."

These observations lead us to recommend that NAEP revise the exclusion guidelines to specify more concrete criteria for inclusion. NAEP should also consider the advisability of providing its own estimation of the functional reading level required for meaningful participation.

Table 16. Factors Cited as Having Had Primary Influence on TSA Inclusion/Exclusion Decisions

	Included in TSA N = 200	Excluded from TSA N = 155	All Students N = 355
Reads/doesn't read well enough to take NAEP	36%	26%	31%
Reading/not reading at grade level	8%	27%	16%
IEP specifies that student should/should not participate	12%	25%	17%
Included/not included in state or district assessment	12%	1%	7%
Could/could not sustain attention for length of assessment	3%	6%	5%
Assessment would/would not be psychologically stressful	7%	3%	5%
Parents would/would not want student to participate	2%	0%	1%
Student did/did not want to participate	3%	0%	1%
Familiar/not familiar with standardized test format	1%	1%	1%
Understands/doesn't understand oral English well enough	3%	0%	1%
Would/would not be disruptive, distract other students	1%	4%	2%
Severe sensory or motor impairment	0%	1%	1%
Severe cognitive or mental impairment	0%	4%	2%
Other (primarily those who said they were unaware that they had the choice to exclude)	15%	3%	10%

Note: Information was gathered from up to two decision makers per student, depending upon which interview respondents reported having been involved in the inclusion/exclusion decision. This table is based on one decision maker per student, with priority given to teacher respondents. 66% of table responses are from teachers, and the remaining 33% are from assessment administrators.

Factors cited as contributing, but not primary to the inclusion/exclusion decision process include more behaviorally oriented criteria (Table 17), such as being able/not able to sustain attention for the length of the assessment, being disruptive/not disruptive for other students, and being placed at risk/not at risk for psychological stress by the assessment.

Table 17. Other Factors Most Frequently Cited as Contributing to TSA Inclusion/Exclusion Decisions

	Included in TSA N = 200	Excluded from TSA N = 155	All Students N = 355
Reads/doesn't read well enough to take NAEP	12%	32%	21%
Reading/not reading at grade level	12%	32%	21%
Could/could not sustain attention for length of assessment	31%	42%	36%
Assessment would/would not be psychologically stressful	16%	25%	20%
Would/would not be disruptive, distract other students	23%	16%	20%

Note: Information was gathered from up to two decision makers per student, depending upon which interview respondents reported having been involved in the inclusion/exclusion decision. This table is based on one decision maker per student, with priority given to teacher respondents. 66% of table responses are from teachers, and the remaining 33% are from assessment administrators.

In order to evaluate the extent to which the primary factors cited as influencing the inclusion/exclusion decisions were really driving the decisions (as opposed to simply being viewed by the respondents as the "correct" response), we also examined the empirical relationship between various student characteristics and the likelihood of inclusion, using both univariate and multivariate analyses. The student characteristics most highly correlated with inclusion are shown in Figure 6. These results are encouraging because they do, after all, show that reading level and percent time mainstreamed are most highly correlated with inclusion. (The latter, of course, is strongly stressed in the official NAEP criteria.) The dominance of these factors was also borne out by a multivariate logistic regression.

Figure 6. Variables Examined for Correlation with TSA Inclusion

Prob of R=0 ≤ .01

	R
Percent time mainstreamed	.56
Reading grade level (teacher report)	.50
Reading instruction in mainstreamed class	.43
Reading grade level (Woodcock Johnson Broad Rding Cluster)	.40
Reading instruction in special education class	-.38
Primary disability = speech impairment	.33
Participated in state, district, or other grade-level stdzed test	.26
Primary disability = mental retardation	-.25
Reported severity of disability (direction of coding: 1 = profound)	.23
Teacher respondent participated in inclusion/exclusion decision	-.16
Assessment participation addressed in IEP	-.14
Race = white	.13
Sensory or orthopedic disability reported	-.12

Prob of R=0 > .01

Number of disabilities reported
Number of years with active IEP plan
Limited English Proficient
Absent more than most students
More disruptive than most students
Primary disability = emotional disturbance
Primary disability = learning disability
Primary disability = sensory impairment
Primary disability = other
AA respondent participated in inclusion/exclusion decision
Sex = male

It therefore seems most probable that the teachers and assessment administrators are abiding, to the best of their ability, with the spirit of the NAEP guidelines, but that, at the same time, they have a conservative view of what level of functioning is necessary for students to

participate "meaningfully" in the assessment.

Research Question 4. Were there reading levels at which a child was likely to be included in some states but excluded in others?

Three-hundred-fourteen students from our study were included in the analysis for the fourth research question on the comparability of exclusions across states. Not included were students without WJ BRC scores, students who had non-cognitive disabilities (e.g., sensory disabilities) that might have influenced the original exclusion decision, and a few students whose data were not cleaned and available at the time these analyses were conducted. Of the 314 students in the analyses, 105 were excluded from the 1994 TSA; Table 18 shows the breakdown of the sample by state.

Table 18. Counts of IEP Students in State Comparison Analysis

State	Excluded from 1994 TSA*	Included in 1994 TSA
1	30	51
2	31	64
3	24	46
4	20	48

*Includes 14 students who had been selected for inclusion in the TSA but did not actually participate because they were absent on the day of testing.

The distributions of grade-equivalent WJ BRC reading scores for the study sample are summarized in Table 19. As can be seen from this table, although all four states tended to exclude poorer readers more often than better readers, there was also substantial overlap in the distributions in each state. Virtually without exception, based on the Woodcock-Johnson test, readers in the top quartile of excluded IEP students read at a higher grade level than readers in the bottom quartile of the included IEP students.

Table 19. Distributions of Reading Grade Levels (WJ BRC grade equivalent scores) for Students in State Comparison Analysis

State x Status	Minimum	1st Quartile	Mean	3rd Quartile	Maximum
1 Included	2.00	2.70	3.55	3.70	7.55
1 Excluded	0.05	1.75	2.45	3.05	5.10
2 Included	1.55	2.95	4.17	5.00	9.40
2 Excluded	0.45	2.08	2.81	3.30	6.70
3 Included	1.45	3.30	4.45	5.65	10.20
3 Excluded	0.25	1.75	2.65	3.1	5.65
4 Included	1.70	2.50	4.07	5.25	8.60
4 Excluded	1.10	2.20	2.81	3.35	4.60

As indicated, the question which we wished to address in this portion of the study is not whether different states excluded different proportions of students or whether they excluded different proportions of students identified with IEPs, but whether they excluded different proportions of IEP students *among those with a particular reading level*. The critical statistic for this analysis, therefore, is where the reading level cutpoint occurs between included and excluded students in each state.

Because there is such great overlap, the estimation of the cutpoint must be statistical. Furthermore, the statistical method used must not be biased by any tendency for some states to categorize either more students or students at higher proficiency levels as having an IEP. (Differences in the latter are influenced by a variety of factors not directly related to our study question, including states' varying ability to allocate resources to special education for children with disabilities.)

The method of analysis was to fit a logistic curve to the data in each state, relating the WJ BRC scores to inclusion.[22] The cutpoint was defined as the grade-level at which the logistic curve crossed .50. For reading at grade levels above this point, students were more likely to be included than excluded; for reading at grade levels below this point, students were more likely to be excluded than included. This point, it should be noted, is also the point of greatest change from exclusion to inclusion as an outcome.

---

[22] Although not essential for this analysis, it is important to note that the WJ BRC correlates highly with NAEP reading scores. As described under the presentation of research question 1, the correlation in this study, based only on included IEP students, was .75. In a less restricted range of students, the correlation is likely to be even higher.

The results are contained in Table 20. The statistical significance for the WJ BRC score was <.01 in each of the four states, and the relationships (coefficients) between variation in WJ BRC scores and variation in inclusion were remarkably similar across states. However, the "intercepts" were different in different states; and some of the differences between states were also statistically significant. In particular, state 1 appears to have had a higher inclusion threshold than the other states, while state 2 had a lower inclusion threshold than the other two states.

Table 20. Relationship of WJ BRC Reading Level to Exclusion from the 1994 TSA

State	WJ coefficient	Intercept	Cutpoint	Significance of between-state differences
1	0.879	3.128	3.56	.003
2	0.724	0.855	0.85	.000
3	0.818	2.417	2.95	.688 (ns)
4	0.786	2.634	3.35	.078 (ns)

The SAS plots of results for the four states are contained in Figures 7 through 10. The first plot in each case is between the WJ BRC score and the logistic probability of inclusion. That plot is necessarily a smooth curve and does show variability in predictiveness. The second plot in each case shows the empirical frequency of WJ BRC scores among included and excluded IEP students, respectively.

Figure 7. Relationship of WJ BRC Reading Level to Inclusion: State 1

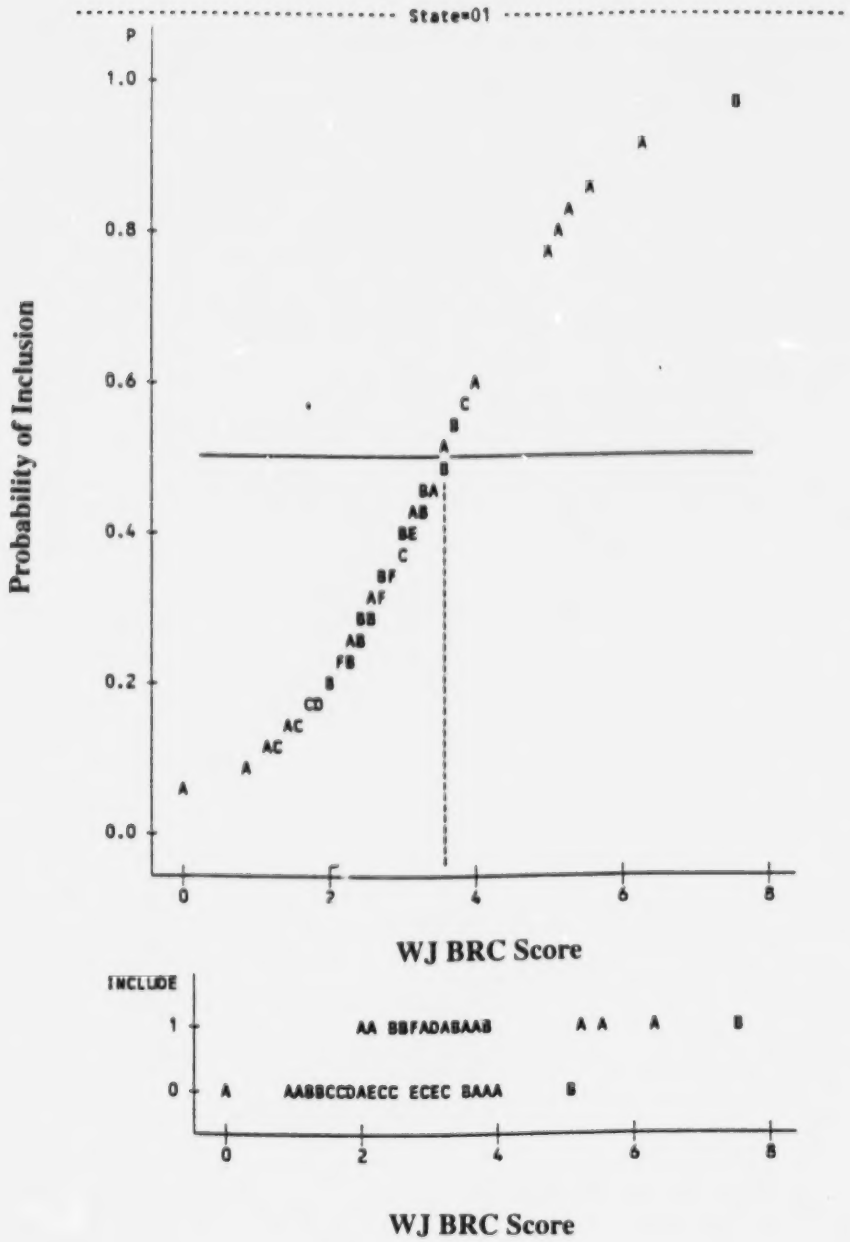


Figure 8. Relationship of WJ BRC Reading Level to Inclusion: State 2

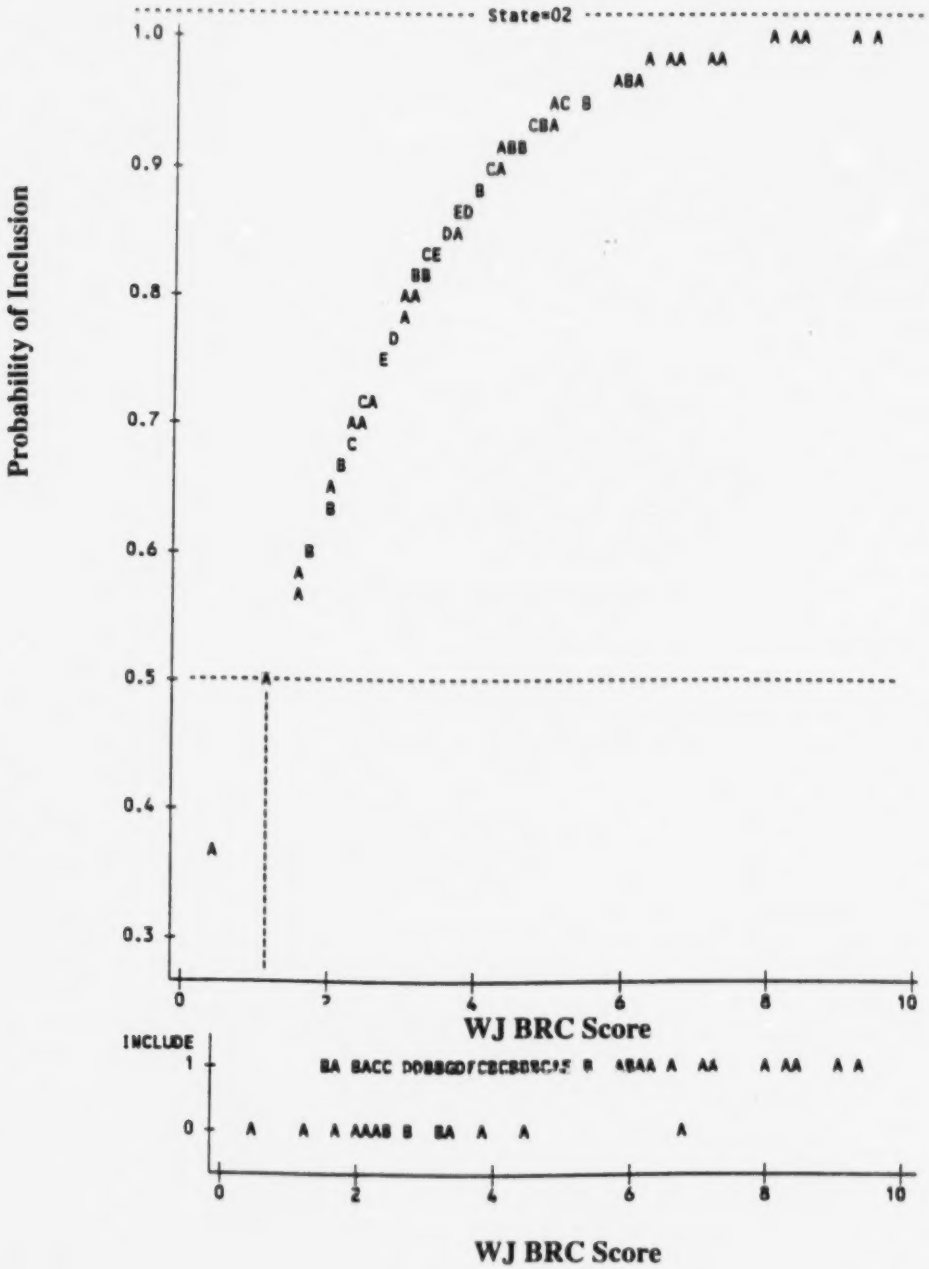


Figure 9. Relationship of WJ BRC Reading Level to Inclusion: State 3

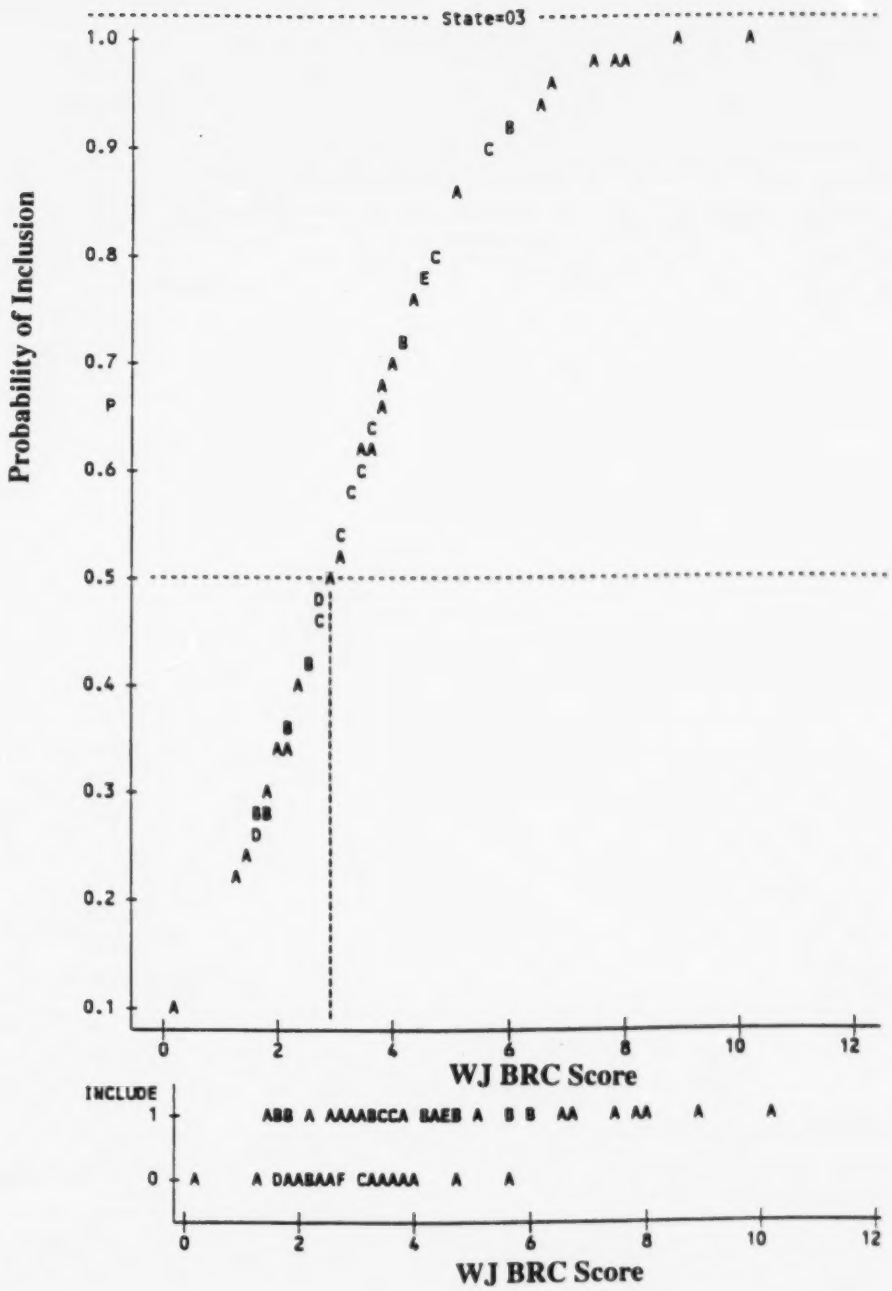
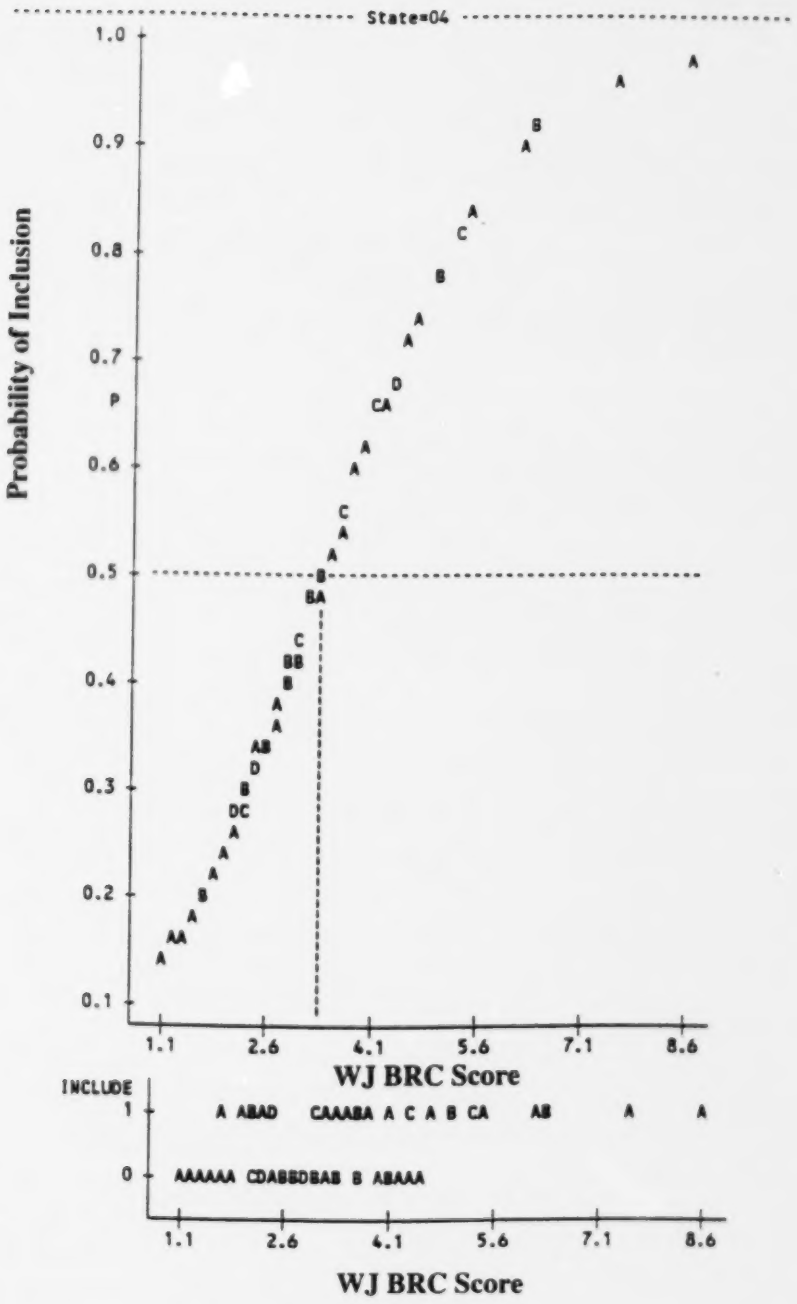


Figure 10. Relationship of WJ BRC Reading Level to Inclusion: State 4



The conclusion from this analysis is that, for whatever reason, there was significant variation between states in the reading level cutpoints for exclusion of IEP students from the 1994 TSA. If these four states' scores had been based on a uniform cutpoint, the state mean reported for state 1 would have been somewhat lower, and the state mean reported for state 2 would have been somewhat higher. To assess the size of this effect, one needs an estimate of the distribution of WJ BRC scores and TSA scores for the excluded IEP students in each state. Then, assuming that a common threshold were set on the Woodcock-Johnson across all four states, imputed mean TSA scores for the states could be generated and compared to the means actually reported.

Although only a sample of IEP students in the four states were included in our study, *IEP/LEP Student Questionnaire* data were available for all the IEP students who had been selected for the TSA sample. In the four study states, these questionnaire data were available for 1,292 students, including 679 who participated in the TSA, 46 who had been absent from the assessment, and 567 who had been excluded. WJ BRC scores were obtained for 337. As a result, an estimate of the reading levels of excluded students and an estimate of the TSA scores for excluded IEP students could be obtained, using the *IEP/LEP Student Questionnaire* data. Two regressions had to be estimated: 1) the WJ BRC scores had to be predicted from the questionnaire data; and 2) TSA scores (mean plausible values) had to be predicted from the questionnaire data (using either the actual WJ BRC scores—for students for whom they were collected—or the estimated WJ BRC scores based on the first equation). The first regression was based on the study sample of 314 cases, augmented to 337 to include cases with all disabilities; and the second regression was based on all included IEP students. All four states were combined for this prediction, so that the resulting prediction equations would be uniform across the states. The value of R^2 for the first regression was .44; the value for the second was .31. The regression functions are shown in Appendix B.

The results are shown in Table 21. The first column of Table 21 displays mean composite plausible values based on participating students; the second column adds to these the estimated scores for all excluded IEP students. The third column deletes from the second column all those cases (included or not) for whom the WJ BRC reading grade level was estimated to be less than 2.0. Each of the first three columns in Table 21 focuses on the states as a whole; the final three columns focus specifically on the IEP students in those states. The excluded students (column 6) were estimated to score lower than the included students (column 4 or column 5) in each state. The differences between columns 4 and 5 reflect the random error of prediction of the plausible values.

In standard analyses of NAEP, excluded students are not allocated cases weights. Therefore, for the purpose of this analysis, all participants were weighted equally. As a result, the state means reported in Table 18 are not exactly the same as published means for those states, although they are very close. As a "bottom line," the largest difference in distances among states

between the first and third columns in Table 21 (the simple mean scores versus the estimated mean scores if a common reading grade level cutoff of 2.0 were used) is the distance between states 2 and 3, which on the simple means are 16.7 points apart and on the estimated means are 18.3 points apart.

Table 21. Unweighted State Average Plausible Values
Under Different Conditions of Inclusion

	All included students who participated	... plus all excluded IEP students	... but only including IEP students with WJ BRC estimate >=2.0	All included IEP students	All included IEP students (estimated)	All excluded IEP students (estimated)
	(n=11,071)	(n=11,638)	(n=11,550)	(n=679)	(n=679)	(n=567)
1	204.5	201.1	202.4	163.2	170.5	152.5
2	221.9	219.5	220.4	179.5	185.4	166.7
3	205.2	201.1	202.1	174.9	174.1	152.0
4	217.6	214.7	215.3	202.6	183.0	162.4

Conclusions and Recommendations

Among the list of principles which the NAE Panel on the Evaluation of the NAEP TSA promulgated in *The Trial State Assessment: Prospects and Realities*, its third report to Congress, was the *Inclusiveness Principle*. This principle states that "To the degree technically, ethically, and financially possible, NAEP should assess an inclusive sample of all children whose ages would ordinarily place them in the 4th, 8th, or 12th grade."[23] Recent trends in educational practices for students with disabilities, as well as legal interpretations that view assessment participation as a civil right, have also combined to increase the saliency of questions of assessibility and exclusion for the NAEP program. By providing evidence on the impact of 1994 TSA exclusion practices and the feasibility of greater inclusion, this study contributed to the deliberations on how to increase inclusion in NAEP, starting with the 1996 assessment cycle.

Assessibility. Based on the observed relationship between students' NAEP scores and scores on the WJ BRC, *we conclude that the large majority of 4th-grade students with disabilities are assessable on the current NAEP instrument if the goal is to achieve a level of measurement that would allow information about these students to contribute to estimates of states' overall performance*. In our study sample, which was representative of the students with disabilities (attending regular schools) in four TSA states, we estimated that 83% were assessable, including 93% of those who had been assessed in 1994, and 70% of those who had been excluded. Our estimate, although derived on an entirely different basis, is remarkably similar to the estimate of 85% assessable reported by the National Center on Educational Outcomes in 1994.

Our data also indicate, however, that the 1994 NAEP reading assessment was not a particularly good fit to the reading proficiencies of IEP students, many of whom are reading a grade or more below grade level. A more appropriate assessment would address exactly the same types of reading behaviors as those expected of other students (and as specified in the NAEP *Reading Framework*), but would utilize easier reading passages and items.

We therefore recommend that NAEP continue efforts to encourage greater participation of students with disabilities on the current assessment, and that results for students with disabilities assessed under standard conditions be aggregated with results for all other students in producing the overall and subgroup achievement estimates normally reported for the nation and the states. NAEP should also work to develop assessments that can measure accurately over a broader range of student achievement levels and thereby provide better estimates at both ends of the achievement distribution..

Accommodations. Our interviews with teachers indicated that they are inclined to be quite liberal in recommending accommodations (or exclusions) for students with disabilities.

---

[23]The National Academy of Education, *The Trial State Assessment: Prospects and Realities* (Stanford, CA: Author, 1993), 96.

BEST COPY AVAILABLE

This may be due to the fact that teachers are most used to making decisions for instructional settings, where the goal would be to facilitate the student's learning in every possible manner. They are less used to weighing the needs for standardization in large scale assessments. In any event, the teachers recommended assessment accommodations for 53% of the students in our sample, including 43% of excluded students and 62% of those who had participated in the 1994 TSA. They also recommended excluding altogether a small percentage of those who had participated, and slightly more than half of those who had not. *Had the teachers' recommendations been followed, only 20% of the 416 IEP students in our study would have been assessed without accommodations in the 1994 TSA, as opposed to the 56% that actually were assessed.*[24] This finding is consistent with the experience of the 1995 NAEP field test, which also found that accommodations were given to a significant percentage of students who otherwise would have been included without accommodation.

Among the learning disabled students who accounted for 67% of our sample, the most commonly suggested accommodations were:

- oral response with or without assistance and interpretation (56%),
- testing in small groups or special education class (70%),
- oral reading of directions with or without interpretation (74%), and
- shorter tests and/or more time for testing (85%).

There was a great deal of diversity in the specific accommodations recommended—for example shortening the test versus extending the testing time versus allowing breaks in testing. *The majority of the recommended accommodations, however, appeared to fall into a few broader categories, and the apparent diversity may simply arise from the fact that there has been little effort (and perhaps little need) to standardize accommodations across students and sites.* The exception would be accommodations for students with sensory or orthopedic disabilities, many of whom would require very specific accommodations adapted to their particular circumstances.

If accommodations are to be offered, their impact on test performance must be understood before the scores of accommodated students can be combined with the scores of other test takers. Our study was not designed to address this question, and we are able to offer only limited insight into the problem. We did estimate the item p-values for the five items in the abridged item block based on the NAEP theta estimates for students in our sample who also took the 1994 TSA. *Comparing these p-values to the observed p-values under the assessment conditions of our study (individual administration, shortened test), we did not observe any pronounced, systematic effect of accommodations.* However, the p-value estimates are not precise because they had to be made using 1992 item parameters with 1994 theta estimates, given

[24] It is only fair to qualify this by noting that there may have been something about the context of the questions in our interview that encouraged the teachers to opt for accommodation. Certainly we never provided them with any arguments concerning the potential problems with score interpretation that could arise from accommodation.

that we had chosen to use a NAEP item set that was released after the 1992 administration and consequently had no 1994 item parameters. The analyses to be undertaken by the NAEP contractor on the results of the 1996 assessment, although in different subject areas (mathematics and science), will provide much better estimates of the impact of accommodations.

We recommend that NAEP continue its present efforts to increase participation of students with disabilities by offering accommodations. However, we also suggest that guidelines for the use of accommodations be developed, with the goal of accommodating only those students for whom the standard NAEP administration would be clearly inappropriate. Further, we recommend that research into the impact of accommodations on performance be continued and that consideration be given to limiting the choice of accommodations to the fewest standardized alternatives that can adequately address the special needs of these students.

Exclusion process. The teachers and other local school personnel who were responsible for implementing the TSA exclusion process did appear to be following (to the best of their ability) the written guidelines provided by NAEP. *The single factor most frequently cited as influencing the inclusion/exclusion decision was the child's reading level.* Choosing from a pre-coded list of options, decision makers for 31% of the children cited "reads/doesn't read well enough to take the NAEP" as the primary influence on their decision; an additional 16% choose the slightly different option, "reading/not reading at grade level."[25] *Further, percent time mainstreamed and teacher's report of the student's functional grade level in reading were the only significant predictors of exclusion status in a multivariate logistic regression that included a number of other factors hypothesized to influence exclusion.* The latter included such factors as student race/ethnicity, student gender, student discipline problems, the instructions regarding assessment in the student's IEP, and the positions of those who participated in the exclusion decision process. The greatest problem appears to be the decision makers' interpretation of the reading achievement level required to participate meaningfully in the assessment.

We recommend that NAEP revise the exclusion guidelines to specify more concrete criteria for inclusion. NAEP should also consider the advisability of providing its own estimation of the functional reading level required for meaningful participation.

Comparability between states. When the exclusion cutpoint is defined as the reading grade level at which the probability of exclusion reaches 50%, we found that *the states selected for our study did differ significantly in their criterion for exclusion decisions. State 1 had the highest cutpoint for exclusion, at an estimated reading grade level of 3.56. State 2 had the lowest cutpoint, at an estimated reading grade level of 0.85. That means that a substantial number of students who would have be included by State 2 would not have been included by State 1.* The implied exclusion criterion of States 3 and 4, with cutpoints of 2.95 and 3.35 respectively, were

---

[25] The NAEP guidelines actually state that the child should be excluded if he or she is determined to be "incapable of participating meaningfully in the assessment;" a fair proportion of respondents apparently operationalized this as "not reading at grade level."

BEST COPY AVAILABL

not significantly different from one another.

These finding add force to our recommendation that exclusion guidelines be modified to specify more concrete criteria for exclusion.

APPENDIX A

Calculating the Criterion for Assessibility

For students on whom both WJ BRC scores and plausible values were available (i.e., students in our sample who had participated in the 1994 TSA) an analysis of the relationship between these two measures of reading achievement was conducted; each student entered the analysis five times, once for each of his or her plausible values. First, a scatter plot of WJ BRC scores by plausible values was constructed and examined visually. (See Figure 2 in main text.) On this basis, it appeared that a grade-equivalent score of 2.1 was a possible cut point below which students' WJ BRC scores no longer predicted NAEP plausible values.

The variables were then entered into a series of linear regressions. In a univariate regression predicting plausible values, the slope of the regression line was very different if the analysis was restricted to students whose WJ BRC scores were at or above 2.1 or to those for whom the score was below 2.1. At or above the cut point, the prediction was significant and the parameter estimate for the regression slope was 15.7. Below the cut point the prediction was not significant and the regression slope turned negative (-25.5).

As a further test of the cut point, a linear regression model was fitted with three independent variables: the WJ BRC score, a dummy variable which indicated whether the student's WJ BRC score was above or below 2.1, and an interaction term between the dummy variable and the WJ BRC score. Over all students in the sample, the adjusted R^2 for this model was .46, as it also was when the model included just the WJ BRC score. In the three variable model, however, only the interaction term was significant, indicating that the relationship between the WJ BRC scores and the plausible values changed at 2.1.

Finally, the F-value for the test of the joint hypothesis that both the slope and the intercept differ on the two sides of the 2.1 cut point was highly significant ($F = 8.6$, $p>F = 0.0002$). When the analysis was rerun, substituting cut points slightly above or below 2.1, the F-value for this test dropped substantial to 5.3 ($p>F = 0.0050$) for cut points slightly above 2.1 (2.2 or 2.3), and even more sharply to 3.3 ($p>F = 0.0386$) and then to a non-significant 0.6 ($p>F = 0.5285$) for cutpoints slightly below (2.0 or 1.9).

Thus, for this group of students, a WJ BRC grade-equivalent score of 2.1 clearly provided the best demarcation between students for whom the 1994 TSA had produced some estimate of their relative reading proficiency (compared to other students in roughly the same proficiency range) and those for whom it had not. However, because the analysis included only 13 students with WJ BRC scores below 2.1, this exact cut score might not generalize to other samples of students. Rather, 2.1, indicating a reading level very near the beginning of grade two, should be taken as an approximation of where the true cut score might lie.

Appendix B

Regression Functions for Predicting WJ BRC Scores and TSA Scores

WJ BRC Score=3.323 + 0.746* (Reading Grade Level, as rated by teacher)
+ 0.366* (Student Included)
- 0.153 * (Age)
- 0.157 * (Sex)
- 0.404 * (Minority)
- 0.073 * (Severity of Disability, as rated by teacher)
+ 1.011* (Percent Time Mainstreamed)
- 0.376 * (Centered Race by Reading Grade Level Interaction)
+ 0.268* (Centered Inclusion by Reading Grade Level Interaction)

TSA Score =120.596 + 13.876 * (WJ BRC)
+ 0.442 * (Reading Grade Level, as rated by teacher)
- 0.296 * (Age)
+ 4.595 * (Sex)
- 11.363 * (Minority)
- 1.631 * (Severity of Disability, as rated by teacher)
+ 21.999 * (Percent Time Mainstreamed)
+ 6.843 * (Centered Race by Reading Grade Level Interaction)

where:

Severity of Disability is coded as 1 to 4, with 4=mild and 1=profound;
Sex=1 for males, =2 for females;
Minority=1 for whites, =2 for all others;
Mainstreamed=0 if less than 50 percent; =.5 if 50 percent to 80 percent; =1 otherwise.

In the equation for TSA scores, actual values of the WJ BRC scores were used where available, and the estimate from the preceding equation was used for other cases.